EE 131A
Probability and Statistics
Instructor: Lara Dolecek

Class Project
Wednesday, 19th July, 2023
TA: Jayanth Shreekumar

Please upload your project to Gradescope by 23rd August 2023 at 11:59 PM PST.
Please submit a single PDF directly on Gradescope.
You may type your project report or scan your handwritten version. Make sure all the work is discernible.
All code shall be submitted in a single zip file on Canvas.
100 points total

In this project, we will further analyze random variables and learn about their utility in practical systems. Each part will have a combination of programming, mathematical analysis, and technical writing. You will be graded on all components.

When producing your plots **clearly indicate** the x-axis, the y-axis, and what is being plotted (using legends, title, etc.). You may need to rescale the x-axis to ensure that your plot is showing the right quantity.

Make sure to attach in the appendix of your project report **all programs/code** that you used to generate the data and plots.

There is no collaboration for the project. You may discuss with other students about the project, but all writing and code must be your own work. The project report and code will be checked for plagiarism.

1. (25 pts) *Tossing a fair and unfair die.* Suppose you have a 12-sided die, with sides numbered 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12.

   (a) Write a program (in any programming language) to simulate the tossing of a 12-sided fair die (i.e., all sides are equally likely), for t = 50, 100, 1000, 2000, 3000, 10000, and 100000 tosses. Based on the simulation, what is the estimated probability of obtaining an odd number?

   (b) Suppose X is a random variable denoting the outcome of a die toss. Based on mathematical analysis and probability theory, what is the probability that X has an odd value?

   (c) Refer back to part (a). Does it agree with the theoretical result in (b)?

   (d) Repeat parts (a), (b), and (c) with a 12-sided die that has the property that any prime number is twice as likely as a non-prime number and all non-prime numbers are equally likely (1 is not a prime number).

2. (25 pts) *Maximum Likelihood Estimation.* Consider the data sequence $\{x_1, x_2, ..., x_n\}$ of length $n = 100000$ provided in the file `'data.txt'`. It is known that the data is normally distributed. However, the mean μ and variance σ^2 of this normal distribution is unknown. In this problem, you are required to estimate μ and σ using methods learned in the course. Assume that the data points are i.i.d. random variables $X_1, X_2, X_3, ..., X_n$ where each $X_i \sim \mathcal{N}(\mu, \sigma^2)$. The data set $\{x_1, x_2, ..., x_n\}$ provided in 'data.txt' are the observed values of these random variables. You will determine μ and σ for the data so that the data matches to its most likely Gaussian bell curve. Mathematically, you determine μ and σ that maximize the probability of observing the data given the choice of μ and σ. Let $f_{X_1,X_2,...,Xn}(x_1, x_2, x_3, ..., x_n|\mu, \sigma)$ be the joint density of the RVs $X_1, X_2, X_3, ..., X_n$ evaluated at our data points $x_1, x_2, ..., x_n$ given our choice of μ and σ. Note that $f_{X_1,X_2,...,X_n}(x_1, x_2, x_3, ..., x_n|\mu, \sigma)$ is a function of μ and σ. The maximum likelihood estimates μ_{MLE} and σ_{MLE} for μ and σ are found as follows:

$$\mu_{MLE} = \arg\max_{\mu} \left[\log(f_{X_1,X_2,...,X_n}(x_1, x_2, x_3, ..., x_n|\mu, \sigma))\right]$$

$$\sigma_{MLE} = \arg\max_{\sigma} \left[\log(f_{X_1,X_2,...,X_n}(x_1, x_2, x_3, ..., x_n|\mu_{MLE}, \sigma))\right]$$

where log is taken to the base e.

(a) Analytically show that

$$\log(f_{X_1,X_2,...,X_n}(x_1, x_2, x_3, ..., x_n|\mu, \sigma)) = -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$$

(b) Find analytical expressions for μ_{MLE} and σ_{MLE} in terms of the data $x_1, x_2, ..., x_n$ by using the equation derived in (a) in the equations for μ_{MLE} and σ_{MLE}. For the data provided in `'data.txt'`, find the values of μ_{MLE} and σ_{MLE} using the analytical expressions derived.

(c) Plot the PDF of X_i by plotting the histogram of the data $\{x_1, x_2, ..., x_n\}$ in `'data.txt'`. For the μ_{MLE} and σ_{MLE} obtained in (b), superimpose the histogram with the PDF of a Gaussian random variable with mean μ_{MLE} and variance σ_{MLE}.

For MATLAB, you can use the function `histogram(data,'Normalization','pdf')` to plot the histogram of the data to represent its PDF.

3. (25 pts) *Naïve Bayes Classifier.* To maximize profit, advertisement companies like Google, Facebook, and Amazon want to target their advertisement to each user. Specifically, given a user and a product to sell, they want to determine whether this user will buy it. In this problem, we will build a classifier to help us determine whether a user will buy a certain product by using historical data.

   Some demographic data is collected for 887 users and are provided in `user_data.csv`. This dataset has four fields: i) Whether this user bought the product (1 for did buy and 0 for did not buy), ii) Type of Spender (1 for larger spender, 2 for medium spender, 3 for small spender) iii) Sex of user (1 for Male and 0 for Female), iv) Age of user. We use random variables B, T, S, and A for bought status, type of spender, sex, and age, respectively. In this problem, we are going to build a popular *Naïve Bayes Classifier* to predict B given T, S, and A.

   (a) Estimate the individual PMFs for B, T, S, and A by finding the fraction of each realization of these random variables among all data. Plot these PMFs (i.e. 4 PMFs in total).

   (b) We are interested in how T, S, and A affect B, in the context of the *Naïve Bayes Classifier*; the first step is to estimate the conditional PMF conditioning on the outcome of interest, i.e., survival or not. Estimate and plot the conditional PMFs for T, S, and A separately conditioned on $B = 1$ and $B = 0$ (i.e. 6 PMFs in total).

   (c) In the *Naïve Bayes Classifier*, we use the conditional independence assumption. For example, if T, S, and A are conditionally independent on B, then $P(T, S, A|B = 0) = P(T|B = 0)P(S|B = 0)P(A|B = 0)$ and $P(T, S, A|B = 1) = P(T|B = 1)P(S|B = 1)P(A|B = 1)$. Using this assumption, compute $P(B = 0, T = 1, S = 0, A \leq 67)$ and $P(B = 1, T = 1, S = 0, A \leq 67)$ based on your estimations in (a) and (b).

   (d) Assuming the conditional independence assumption and using your result in (c), compute $P(B = 0|T = 1, S = 0, A \leq 67)$ and $P(B = 1|T = 1, S = 0, A \leq 67)$. Predict whether a female whose age is below 67 and who is a large spender will buy this product or not.

4. (25 pts) *Central Limit Theorem.* Let X_1, X_2, X_3, ... be a sequence of i.i.d. random variables with finite mean μ and finite variance σ^2, and let Z_n be the mean of the first n random variables in the sequence:

$$Z_n = \frac{X_1 + X_2 + ... + X_n}{n}.$$

In this problem, we will plot the PDF of Z_n by generating $t = 10^4$ samples of Z_n and plotting the histogram of the generated samples using the MATLAB function `histogram(samples,'Normalization','pdf')`

(a) Let X_i, for $i = 1, 2, \ldots$ be a uniform continuous random variable taking values in the interval $(10, 16)$. Write a program to plot the PDF of Z_n. Consider $n = 1, 2, 3, 10, 30, 100$ and compare your results across different n's.

(b) Calculate analytically the mean and the variance of X_i and of Z_n in part (a).

(c) Write a program to generate a Gaussian random variable with the same mean and variance as Z_n. Superimpose its pdf on the plots from part (a).

(d) Repeat parts (a), (b), and (c) with X_i representing a toss of a 12-sided die that is described in Problem 1(d). Note that X_i and Z_n are discrete in this case. Hence, we approximate the PDF of Z_n by generating samples of Z_n and plotting its histogram. To plot the histogram of Z_n use the function `histogram(samples,'Normalization','pdf', 'BinWidth', 1/(n+1) )`.

Remark: In the case when X_i's are discrete (which leads to discrete Z_n), the upper faces of the rectangles forming the histogram of Z_n converges toward a Gaussian PDF as n approaches infinity and hence even in this case, you should be able to observe the Central Limit Theorem.

(e) *(Bonus 10 points)* Repeat parts (a) and (c) with the same X_i described in part (a), but this time compute and plot the exact PDF of Z_n instead of generating samples. Provide full description of how you computed the exact PDF of Z_n.